

OVSegDT: Segmenting Transformer for Open-Vocabulary Object Goal Navigation

Tatiana Zemskova^{1 2}, Aleksei Staroverov¹, Dmitry Yudin^{1 2}, Aleksandr Panov^{1 2}

¹AIRI, 32 Kutuzovsky Ave., Moscow, 121170, Russia

²Moscow Institute of Physics and Technology, 9 Institutsky per., Dolgoprudny, 141701, Russia

Abstract

Open-vocabulary Object Goal Navigation requires an embodied agent to reach objects described by free-form language, including categories never seen during training. Existing end-to-end policies overfit small simulator datasets, achieving high success on training scenes but failing to generalize and exhibiting unsafe behaviour (frequent collisions). We introduce OVSegDT, a lightweight transformer policy that tackles these issues with two synergistic components. The first component is the semantic branch, which includes an encoder for the target binary mask and an auxiliary segmentation loss function, grounding the textual goal and providing precise spatial cues. The second component consists of a proposed Entropy-Adaptive Loss Modulation that is a per-sample scheduler that continuously balances imitation and reinforcement signals according to the policy entropy, eliminating brittle manual phase switches. These additions cut the sample complexity of training by 33%, and reduce collision count in two times while keeping inference cost low (130M parameters, RGB-only input). On HM3D-OVON, our model matches the performance on *unseen* categories to that on *seen* ones and establishes state-of-the-art results (40.1% SR, 20.9% SPL on val unseen) without depth, odometry, or large vision–language models. Code is available at <https://github.com/CognitiveAISystems/OVSegDT>.

Introduction

Enabling robots to autonomously navigate toward a target object in an unfamiliar indoor environment is a critical capability for a wide range of real-world applications, including domestic assistance, search and rescue, and human-robot interaction. In this setting, the agent must perceive its surroundings, understand the task instruction, and explore the environment effectively to locate and approach the specified object. This task, commonly referred to as Object Goal Navigation (ObjectNav), presents a number of challenges, particularly when the environment is previously unseen and the object is described using open-ended, natural language.

Early work (Staroverov et al. 2020; Maksymets et al. 2021; Ramrakhya et al. 2023) on ObjectNav typically framed the problem as navigation to one of a fixed set of object categories (e.g., "chair", "table"). While this formulation simplifies the task, it falls short of reflecting the complexity and flexibility of real-world scenarios. A more realistic setup is one where the agent can navigate to an

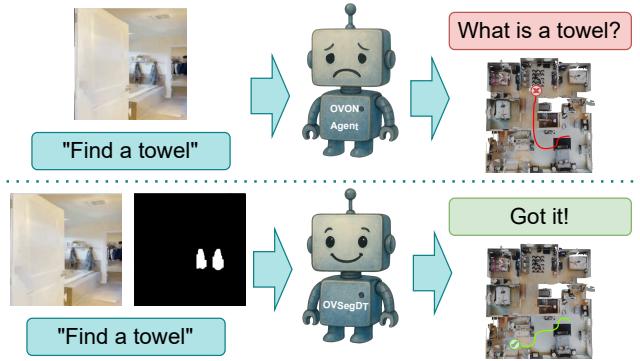


Figure 1: We demonstrate that the explicit integration of semantic segmentation through a target binary mask and an auxiliary semantic segmentation loss function significantly improves the quality of end-to-end training of a transformer-based model for open-vocabulary navigation.

arbitrary object category specified by a natural language query, including categories that were not seen during training (Yokoyama et al. 2024b). This open-ended formulation better aligns with how humans describe goals and greatly increases the generalization ability of the agent.

Generally, the ObjectNav task can be broken down into three stages: exploring the environment, recognizing the target object, and reaching the target object. During exploration, the agent must actively search an unknown environment until it encounters an object that matches the goal description. Upon target recognition, the agent must navigate precisely to the object and stop within a given distance threshold. Existing solutions handle this pipeline either through modular approaches (Staroverov et al. 2020; Yin et al. 2024; Cai et al. 2024), which separate the system into dedicated skills (e.g., environment exploration, goal recognition, and goal reaching), or via end-to-end methods (Maksymets et al. 2021; Zhang et al. 2024; Yokoyama et al. 2024b), where a single model implicitly learns to switch between different stages of the pipeline.

Modular map-based approaches have shown strong performance in open-vocabulary object goal naviga-

tion (Yokoyama et al. 2024a; Yin et al. 2024; Chang et al. 2023), largely due to their ability to leverage large language models to perform decision-making processes based on free-form textual descriptions of the goal and open-vocabulary object detectors to build semantic maps. Once detected, the positions of candidate objects can be projected onto a spatial map, effectively reducing the task to navigation to a given location on the map. Assuming accurate localization, a robot can reliably navigate to a specified point on the map using either a deterministic policy (e.g. (Sethian 1999)) or a learned skill (e.g. (Wijmans et al. 2020)). This structured decomposition simplifies planning and allows the system to benefit from advances in vision-language models and mapping techniques.

However, map-based pipelines are inherently sensitive to the quality of semantic map construction: errors in object detection or pose estimation can propagate and compound through the system, leading to a performance drop in real-world scenarios. End-to-end mapless approaches (Cai et al. 2024; Zhang et al. 2024; Yokoyama et al. 2024b) operate directly on raw sensor inputs without explicit intermediate representations. This can reduce error accumulation and allow the system to learn robust, task-specific strategies that adaptively integrate perception, language understanding, and control. Moreover, such approaches require only a minimal sensor setup and are particularly appealing for real-world deployment where hardware constraints are significant challenges.

End-to-end ObjectNav approaches typically rely on a fragile two-stage learning pipeline: an initial behavior cloning phase (e.g., DAgger) to bootstrap exploration, followed by a hard switch to Proximal Policy Optimization (PPO) for refinement. The transition point between these objectives is chosen heuristically, often leading to catastrophic performance drops when the learning signals conflict (Yokoyama et al. 2024b). To overcome this, we introduce Entropy-Adaptive Loss Modulation (EALM), which dynamically mixes imitation and reinforcement signals at the sample level. By deriving the PPO loss weight from policy entropy in each mini-batch, EALM smoothly shifts emphasis from imitation to on-policy learning without manual scheduling, enabling stable, single-stage training.

A central challenge for end-to-end, mapless ObjectNav in open-vocabulary settings is the availability of photo-realistic training data. Even the largest publicly available simulator dataset, HM3DSem (Yadav et al. 2023), comprises only 216 annotated scenes, which is orders of magnitude smaller than the corpora used to pretrain modern vision–language foundation models. Consequently, learned policies fail to generalize to novel object categories (Yokoyama et al. 2024b). To address this data bottleneck, recent work adopts a hybrid, mapless architecture that decouples exploration from goal-directed navigation: an exploration policy roams freely until an external open-vocabulary detector localizes the target, at which point control is handed off to a specialized navigational skill. This switch can be effected via binary object masks (Cai et al. 2024) or by estimating 3D goal coordinates from depth and odometry measurements (Yokoyama et al. 2024b). These approaches combine the sample effi-

ciency of modular components with the adaptability of end-to-end learning.

In this work, we show that adding a binary mask of the target object to the input of a transformer-based model for open-vocabulary navigation speeds up learning and improves generalization to new object categories (see Figure 1). We also demonstrate that solving an auxiliary semantic segmentation task during training further speeds up policy convergence. Finally, we evaluate our approach in a realistic setup without access to ground-truth scene segmentation, showing that strong generalization to novel categories is still achieved using pretrained open-vocabulary segmentation models. Our method, OVSegDT, achieves state-of-the-art results on the open-vocabulary object goal navigation benchmark HM3D-OVON (Yokoyama et al. 2024b) in a photo-realistic environment.

To summarize, our main contributions are as follows:

- OVSegDT, a transformer-based architecture for open-vocabulary navigation, that incorporates a semantic encoder for binary target masks into the policy’s observation space. This goal mask encoder accelerates policy learning and enables robust navigation to both seen and unseen object categories, even with noisy predicted segmentation.
- Auxiliary segmentation objective that helps disentangle the observation vector in a transformer-based navigation model into components for scene understanding (RGB) and goal representation (mask), improving both convergence and navigation quality.
- Entropy-Adaptive Loss Modulation (EALM), an automatic DAgger–PPO switcher that balances imitation and reinforcement signals without manual scheduling.

Related works

The task of navigating to a target object described in natural language has gained attention as an extension of classic Object Goal Navigation (ObjectNav). Unlike a closed-vocabulary setting where the target object belongs to a pre-defined set of categories, open-vocabulary navigation requires agents to generalize to novel unseen object categories at inference time. In this task, the category names are specified as text, so open-vocabulary object goal navigation falls under the domain of Vision-Language Navigation (VLN).

Map-based open-vocabulary ObjectGoal navigation. Several works incorporate explicit mapping and pre-trained vision-language models to support open-vocabulary understanding. Early works that leverage large language models and explicit mapping, such as NavGPT (Zhou, Hong, and Wu 2024) and MapGPT (Chen et al. 2024), focused on the Room-to-Room navigation task (Anderson et al. 2018). This task requires following language instructions to navigate waypoints, while open-vocabulary object goal navigation involves autonomous search in unknown environments, making it more realistic for real-world robots. Considering methods for open-vocabulary object goal navigation, it is worth noting the VLFMs (Vision-Language Frontier Maps) (Yokoyama et al. 2024a) method, proposing constructing semantic frontier maps that leverage vision-

language models to guide navigation without requiring task-specific training. In parallel, ESC (Zhou et al. 2023) employs soft commonsense constraints during exploration to improve sample efficiency and reduce semantic ambiguity in object-goal navigation. Similarly, SG-Nav (Yin et al. 2024) introduces dynamic 3D scene graph prompting, allowing large language models to reason over structured spatial representations and navigate to unseen objects. TANGO (Ziliootto et al. 2025) proposes a neuro-symbolic, LLM-based framework for embodied AI that leverages primitive modules and extends prior exploration policies to multi-goal settings via memory map mechanisms. While effective, these approaches depend heavily on the quality of the map and robot localization, and errors in either can significantly degrade performance. At the same time, our method does not rely on building a map of the surrounding environment, and thus is independent of the quality of depth and odometry sensors.

Mapless open-vocabulary ObjectGoal navigation. In contrast, mapless methods for open-vocabulary object goal navigation attempt to learn a policy that directly maps observations and language instructions to actions without explicit mapping. These works include, in particular, end-to-end approaches that map observations directly to action predictions. One such method is DagRL (Yokoyama et al. 2024b), which shows that integrating pretrained visual backbones and conditioning the policy on object goal text embeddings enables generalization to unseen categories. PSL (Sun et al. 2024) introduces a training strategy that prioritizes semantic learning, improving zero-shot navigation. However, end-to-end models are typically trained on millions of observations collected in a limited number of environments, such as HM3DSem (Yadav et al. 2023), ProcTHOR (Deitke et al. 2022), and AI2Thor (Kolve et al. 2017), which can hinder generalization to novel goals. To address this, Uni-Navid (Zhang et al. 2024) leverages pretrained language models, while methods like PoliFormer (Zeng et al. 2025) and DagRL-OD (Yokoyama et al. 2024b) incorporate visual cues from open-vocabulary object detectors. In our experiments, we use semantic masks of the target to provide precise location and boundary cues. PixelNav (Cai et al. 2024) also uses such masks but converts them into pixel goals and relies on additional LVLM and LLM modules for global planning. In contrast, we demonstrate that a lightweight transformer model (130M parameters) can efficiently explore environments and leverage open-vocabulary segmentation masks for effective open-vocabulary navigation.

Learning Auxiliary Tasks for Navigation. To address the inherent difficulty of learning robust policies in high-dimensional, multimodal settings, prior work has explored multi-task learning as a way to enrich the agent’s representations and promote generalization. These approaches typically introduce auxiliary tasks that complement the main navigation objective and encourage better grounding of visual and language inputs.

For transformer-based navigation models, there are two common types of auxiliary tasks. The first type includes linguistic tasks such as visual question answering (Uni-Navid (Zhang et al. 2024), NaVILA (Cheng et al. 2024))

and reasoning (Wang et al. 2025b). However, such tasks are typically effective for large transformer models, whereas we use a lightweight state encoder with only 34M parameters. The second type involves purely visual tasks, where the vision-language-action model is pre-trained to generate visual outputs from textual instructions and then fine-tuned to predict actions. Examples include GR-1 (Wu et al. 2024) and 3D-VLA (Zhen et al. 2024), where the model learns to predict the goal’s final state, represented by an RGB image and a depth map. PixelNav (Cai et al. 2024) additionally uses tracking and step-distance to the goal as auxiliary losses. In contrast to these approaches, semantic segmentation combines both visual and semantic scene understanding, which is crucial for navigating toward object categories in indoor environments that often exhibit high visual diversity.

In previous works such as A2CAT-VN (Kulhánek et al. 2019) and MTU3D (Zhu et al. 2025), segmentation was used as an auxiliary task to help form better representations of the current observation. In contrast, we use segmentation as an auxiliary task for constructing the current state representation conditioned on past observations - that is, our transformer model jointly solves both navigation and semantic segmentation tasks. In this sense, our work is related to the Unified-IO2 (Lu et al. 2024) method, which builds a universal instruction-following model capable of both action prediction and image segmentation. However, unlike this 1B-parameter model, OVSegDT is lightweight and can incorporate observation history for navigation tasks, whereas Unified-IO2 is limited to manipulation tasks. Through experiments, we demonstrate that combining an auxiliary segmentation loss with the input goal mask improves navigation performance toward text-specified targets, both for object categories seen during training and for unseen ones. This enables a lightweight and generalizable approach to open-vocabulary object navigation using only an RGB sensor.

Task Setup

We formulate the Open-Vocabulary Object Navigation (OVON) task as a partially observable Markov decision process (POMDP) $\langle \mathcal{S}, \mathcal{A}, P, R, \rho_0, \gamma \rangle$ with state space \mathcal{S} , discrete action space \mathcal{A} , transition distribution P , reward function R , initial state distribution ρ_0 , and discount factor γ .

At every timestep t the agent receives the observation

$$o_t = (I_t, g, a_{t-1}, M_t), \quad (1)$$

where $I_t \in \mathbb{R}^{H \times W \times 3}$ is the current RGB frame, g is the natural-language goal description, $a_{t-1} \in \{0, \dots, 5\}$ is the previous discrete action, and $M_t \in \{0, 1\}^{H \times W}$ is the binary segmentation mask for the target category at the current step.

The action set consists of six high-level motor commands: `stop`, `forward` (0.25 m), `turn_left` (30°) and `turn_right` (30°), `look_up` (30°), and `look_down` (30°). An episode terminates when either `stop` is issued or the horizon of 500 steps is reached. The agent receives a sparse reward of +1 for successfully ending an episode and a small negative reward for collisions.

We adopt the standard ObjectNav metrics – Success Rate (SR), Success weighted by Path Length (SPL), and SoftSPL – following Yokoyama et al. (2024b).

Method

In this section, we first describe the architectural features of our method, and then detail the proposed training objectives that we use to accelerate the convergence of OVSegDT training in the interactive simulator environment. Finally, we discuss how our method can be adapted to a real-world setup where ground-truth target segmentation from the simulator is not available.

Preliminaries

Notation. The transformer policy is denoted by $\pi_\theta(a | s)$ with parameters θ and value head $V_\theta(s)$. Advantage estimates are \hat{A}_t and returns \hat{R}_t .

Proximal Policy Optimisation (PPO) For an on-policy batch $\{(s_t, a_t, \hat{A}_t, \hat{R}_t)\}$ we minimize the clipped surrogate

$$\mathcal{L}_{\text{PPO}}(\theta) = \mathbb{E}_t \left[-\min(r_t \hat{A}_t, \text{clip}(r_t, 1-\varepsilon, 1+\varepsilon) \hat{A}_t) \right], \quad (2)$$

$$r_t = \frac{\pi_\theta(a_t | s_t)}{\pi_{\theta_{\text{old}}}(a_t | s_t)}. \quad (3)$$

A value term $\mathcal{L}_V(\theta) = \frac{1}{2} \|V_\theta(s_t) - \hat{R}_t\|_2^2$ and entropy bonus H_t are added following standard practice.

Dataset Aggregation (Dagger) Given teacher actions a_t^E from a classical A* planner (Hart, Nilsson, and Raphael 1968) that uses a ground-truth map and GPS sensor with a frontier-based exploration strategy for goal object searching (Santosh, Achard, and Jawahar 2008), behavior cloning minimizes

$$\mathcal{L}_{\text{BC}}(\theta) = \mathbb{E}_t [-\log \pi_\theta(a_t^E | s_t)]. \quad (4)$$

Importantly, our implementation also updates the critic during behavior cloning, enabling a seamless transition to RL.

OVSegDT architecture

Following the state-of-the-art architecture for open-vocabulary navigation proposed by the authors of the HM3D-OVON (Yokoyama et al. 2024b) benchmark, we use a frozen SigLIP (Zhai et al. 2023) encoder for RGB images and text, as these encoders have demonstrated strong performance on the ObjectNav task (Ehsani et al. 2023). To encode the observation at time step t , the image encoder $\text{SigLIP}_{\text{RGB}}$ produces a 768-dimensional embedding $i_t = \text{SigLIP}_{\text{RGB}}(I_t)$ for the visual input I_t . Similarly, the text encoder $\text{SigLIP}_{\text{text}}$ generates a 768-dimensional embedding $g_t = \text{SigLIP}_{\text{text}}(G)$ for the object category G provided as text.

We also use a learnable 32-dimensional embedding for the previous discrete action $p_t = \phi_a(a_{t-1})$. To encode the binary target mask M_t , we use a simple, lightweight ResNet9 architecture (He et al. 2016), which has proven efficient in our experiments. This results in a 128-dimensional goal mask embedding $m_t = \text{ResNet9}(M_t)$. The resulting embeddings are concatenated to form the observation embedding $o_t = [i_t, g_t, p_t, m_t]$.

A sequence of the most recent 100 observation embeddings $[o_{t-99}, \dots, o_t]$ is passed to a decoder-only transformer

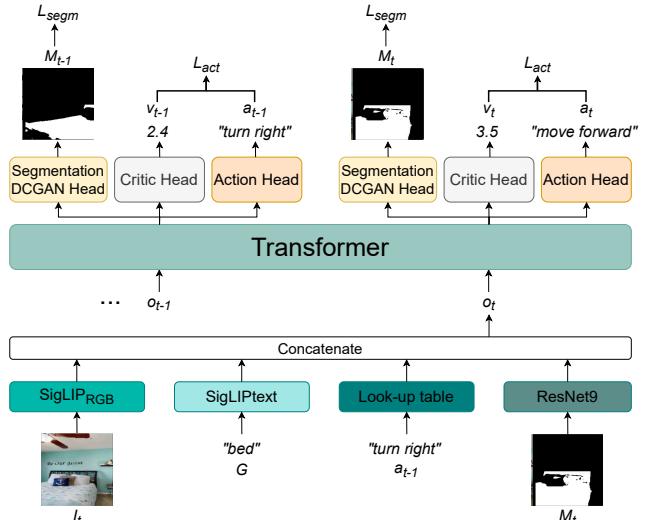


Figure 2: OVSegDT model encodes the current visual observation I_t , the goal object category G , the previous action a_{t-1} and binary mask of the target object (goal mask) M_t to form observation embedding o_t . At each timestep, a transformer receives the embedding sequence from the previous 100 steps. The action head is used to sample the action a_t . During training, a segmentation DCGAN head generates a binary target mask m_t for the auxiliary segmentation loss function, and a critic head is employed to predict the value v_t of the current state.

model π_θ (Vaswani et al. 2017). We adopt the same transformer architecture found effective in the original OVON method: 4 layers, 8 heads, hidden size 512, and a maximum context length of 100. At each timestep, π_θ predicts a feature vector, which is passed to decoding heads.

The action head is implemented as a linear layer that predicts a categorical distribution from which the action a_t is sampled, i.e., $a_t \sim \pi_\theta(\cdot | o_{t-99}, \dots, o_t)$. During training, we also use two additional heads: a critic head, implemented as a linear layer, and a segmentation head, implemented as a DCGAN (Radford, Metz, and Chintala 2015) model that reconstructs the current binary target mask M_t from the feature vector predicted by the transformer π_θ .

Training objectives

In this section, we describe the training objectives used to train our OVSegDT model. Our primary goal was to develop a unified single-stage training pipeline to obtain a general-purpose navigation model capable of using binary target masks as visual cues for navigating to object categories that were not present in the training dataset.

Entropy-Adaptive Loss Modulation (EALM). Open-vocabulary navigation requires a policy to decide how strongly it should rely on imitation signals versus exploration-driven reinforcement learning. EALM solves this by continuously blending Behavior Cloning (BC) and

Proximal Policy Optimization (PPO) objectives according to the policy’s uncertainty, quantified by its action entropy.

Entropy tracking. For every mini-batch we compute the Shannon entropy of the action distribution

$$H_t = -\sum_{a \in \mathcal{A}} \pi_\theta(a | s_t) \log \pi_\theta(a | s_t), \quad (5)$$

and maintain an exponential moving average (EMA): $\hat{H}_t = \alpha \hat{H}_{t-1} + (1 - \alpha) H_t$, $0 < \alpha < 1$. We adopt $\alpha = 0.95$ throughout, a value that works well across domains.

Adaptive gate. Given two entropy bounds $H_{\text{low}} < H_{\text{high}}$ (we use 0.35 and 0.75 in practice), the EMA is mapped to a mixing coefficient

$$\lambda_t = \text{clip}\left(\frac{H_{\text{high}} - \hat{H}_t}{H_{\text{high}} - H_{\text{low}}}, 0, 1\right). \quad (6)$$

Low entropy (confident policy) therefore yields $\lambda_t \approx 1$ (pure PPO), whereas very high entropy favors BC. We further define $p_{\text{PPO}} = \lambda_t$, $p_{\text{BC}} = 1 - \lambda_t$.

EALM objective. With \mathcal{L}_{PPO} the usual clipped surrogate and $\mathcal{L}_{\text{BC}} = -\log \pi_\theta(a_t^E | s_t)$ the imitation loss, the policy term becomes

$$\mathcal{L}_{\text{EALM}}(\theta) = p_{\text{PPO}} \mathcal{L}_{\text{PPO}}(\theta) + p_{\text{BC}} \mathcal{L}_{\text{BC}}(\theta). \quad (7)$$

EALM therefore provides a smooth, automatic transition from demonstration-driven learning to pure reinforcement learning, eliminating the need for manual phase scheduling.

Auxiliary semantic segmentation loss. To encode the observation, we use a single vector obtained by concatenating different types of information extracted from the environment. This helps reduce the time required to predict the next action and enables the model to retain a long observation history, which is particularly important for efficient learning in navigation tasks. However, when working with such a concatenated observation vector, the transformer model must learn how to effectively utilize the information contained in different parts of the observation. In our experiments, we show that using an auxiliary semantic segmentation loss allows the transformer to learn effective navigation representations more quickly during the early stages of training, where the behavior cloning term dominates the navigation training objective. As the loss function for semantic segmentation, we use the sum of the Dice loss and binary cross-entropy loss, which are commonly used as training objectives for binary mask segmentation (e.g., in (Cheng et al. 2022)):

$$\mathcal{L}_{\text{seg}}(\theta) = \mathcal{L}_{\text{Dice}}(\theta) + \mathcal{L}_{\text{CE}}(\theta). \quad (8)$$

Total training loss. Finally, the overall optimization target for a mini-batch is

$$\mathcal{L}_{\text{total}}(\theta) = c_v \mathcal{L}_V(\theta) + L_{\text{EALM}}(\theta) - \beta H_t + \mathcal{L}_{\text{seg}}(\theta), \quad (9)$$

where \mathcal{L}_V is the value regression term and H_t an entropy bonus. All coefficients are kept task-invariant across experiments.

Adaptation to predicted segmentation

Our experiments show that using a binary ground-truth mask of the target not only accelerates the model’s training process but also significantly improves the adaptability of the approach to new environments and target categories. However, in real-world scenarios, ground-truth segmentation is not available, and the quality of navigation heavily depends on the performance of the pre-trained segmentation model used. Given the universal nature of the task and the availability of modern open-vocabulary segmentation models, we aim to leverage a pre-trained segmentation model, as this allows us to minimize the real-to-sim gap for segmentation.

In our experiments, we use a pre-trained YOLOE (Wang et al. 2025a) model for open-vocabulary segmentation – a state-of-the-art model with real-time inference speed. When using this model out of the box, we observe a significant drop in navigation performance compared to using ground-truth segmentation. To mitigate this drop, we apply two strategies. First, YOLOE struggles to recognize masks of objects that are distant or heavily occluded by other objects. To address this issue, we fine-tune our OVSegDT model using pure PPO loss with ground-truth target masks filtered by size. Specifically, we discard masks if their width or height is less than 32 pixels or if the total number of pixels in the mask is fewer than 1000.

Second, we found that different object categories pose varying levels of difficulty for the segmentation model. For example, accurately recognizing a “dishwasher” requires discarding all predicted masks with confidence below 0.4, whereas detecting objects such as “book” or “flowerpot” requires keeping all masks with confidence no lower than 0.01. To address this, we calibrate the detector’s confidence thresholds for different object categories in the model’s navigation vocabulary, which significantly improves navigation quality using predicted segmentation (see Table 4). Finally, we remove semantically redundant categories (e.g., “rug” and “carpet”) from the segmentation vocabulary, keeping only one of each such set. In practice, a confidence threshold of 0.3 is generally robust for segmentation and therefore navigation to unseen categories. The described adaptation of the segmentation model to the specific scenario of robot deployment is a significantly less expensive operation than retraining the navigation model, which takes tens of millions of steps.

Training details

We train all policies for 200M steps, as incremental improvements diminished beyond this point. All experiments were conducted across 40 environments distributed across 2 NVidia A100 GPUs utilizing Variable Experience Rollout (Wijmans, Essa, and Batra 2022). We provide the list of used hyper-parameters and navigation reward details in the Appendix.

Results

Experimental setup. We evaluate the effectiveness of different methods using three metrics: SR, SPL, and the average number of collisions per episode. For methods whose

Method	Depth	Odometry	Val Seen		Val Seen Synonyms		Val Unseen	
			SR↑	SPL↑	SR↑	SPL↑	SR↑	SPL↑
BC (Yokoyama et al. 2024b)	✗	✗	11.1 ± 0.1	4.5 ± 0.1	9.9 ± 0.4	3.8 ± 0.1	5.4 ± 0.1	1.9 ± 0.2
DAGger (Yokoyama et al. 2024b)	✗	✗	18.1 ± 0.4	9.4 ± 0.3	15.0 ± 0.4	7.4 ± 0.3	10.2 ± 0.5	4.7 ± 0.3
RL (Yokoyama et al. 2024b)	✗	✗	39.2 ± 0.4	18.7 ± 0.2	27.8 ± 0.1	11.7 ± 0.2	18.6 ± 0.3	7.5 ± 0.2
BCRL (Yokoyama et al. 2024b)	✗	✗	20.2 ± 0.6	8.2 ± 0.4	15.2 ± 0.1	5.3 ± 0.1	8.0 ± 0.2	2.8 ± 0.1
DAGRL (Yokoyama et al. 2024b)	✗	✗	41.3 ± 0.3	21.2 ± 0.3	29.4 ± 0.3	14.4 ± 0.1	18.3 ± 0.3	7.9 ± 0.1
Uni-NaVid (Zhang et al. 2024)	✗	✗	41.3	21.1	43.9	21.8	39.5	19.8
VLFM (Yokoyama et al. 2024a)	✓	✓	35.2	18.6	32.4	17.3	35.2	19.6
DAGRL+OD (Yokoyama et al. 2024b)	✓	✓	38.5 ± 0.4	21.1 ± 0.4	39.0 ± 0.7	21.4 ± 0.5	37.1 ± 0.2	19.8 ± 0.3
TANGO (Ziliotto et al. 2025)	✓	✓	-	-	-	-	35.5 ± 0.3	19.5 ± 0.3
MTU3D (Zhu et al. 2025)	✓	✓	55.0	23.6	45.0	14.7	40.8	12.1
OVSegDT	✗	✗	41.0 ± 1.0	21.6 ± 0.6	37.7 ± 1.7	19.6 ± 1.7	40.1 ± 1.1	20.9 ± 0.9

Table 1: Open-Vocabulary Object goal navigation. Comparison on HM3D-OVON (Yokoyama et al. 2024b).

navigation relies on sampling actions from distributions, we average the results across three seeds and report the mean and standard deviation. Details of the statistical analysis are provided in the Appendix.

Comparison with the State-of-the-Art methods

We compare our method to state-of-the-art approaches on HM3D-OVON (Yokoyama et al. 2024b), a benchmark tailored for open-vocabulary object-goal navigation. In these experiments, we use the pretrained YOLOE (Wang et al. 2025a) model as the source of goal segmentation. Table 1 shows the navigation performance of our approach compared to existing methods for open-vocabulary navigation. Notably, our method demonstrates improvements in both the success rate and the efficiency in terms of path length (SPL metric) over baseline RGB-only approaches that do not leverage large language models. We provide a trajectory analysis of our method compared to the baseline DagRL, which does not use segmentation masks as input, in Figure 3. These examples illustrate that, during inference, the binary goal mask helps the agent recognize the target object and switch from exploration to goal-directed navigation, while also reducing the distance traveled to reach the target. It is worth noting that our approach outperforms large VLN models such as Uni-NaVid (Zhang et al. 2024) and TANGO (Ziliotto et al. 2025), as well as methods that rely on additional sensors such as depth and odometry for navigation, including DAGRL+OD (Yokoyama et al. 2024b) and MTU3D (Zhu et al. 2025) on the val unseen split of HM3D-OVON. Furthermore, the binary goal mask input makes our approach highly generalizable, as the navigation performance on both val seen and val unseen categories is nearly identical.

Ablation studies

We analyze the effectiveness of individual components of our approach through a series of ablation experiments.

Comparison of Training Strategies. To evaluate our proposed training paradigm, we conduct an extensive comparison of different learning approaches for open-vocabulary object navigation. We analyze four distinct strategies: (1) PPO, (2) DAGger (3) EarlySwitcher that smoothly switches from DAGger to PPO at steps 40M-60M, (4) DAGger+PPO combined loss, (5) DAGRL baseline (Yokoyama et al. 2024b),

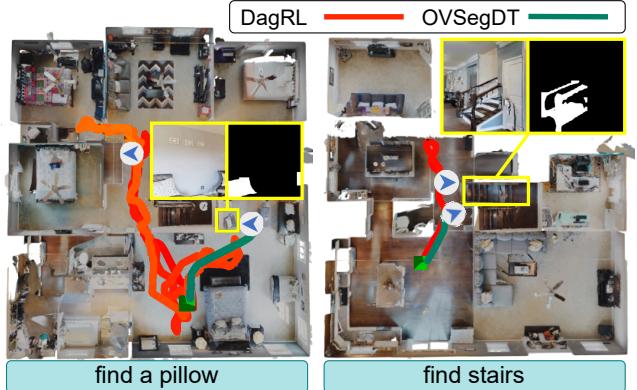


Figure 3: Qualitative comparison of navigation trajectories between our method OVSegDT and the baseline DagRL on categories from the *val unseen* split of the HM3D-OVON benchmark. **Left:** The DagRL model fails to recognize the target object *a pillow* and thus ends the episode unsuccessfully, while OVSegDT successfully reaches the goal thanks to the binary goal mask. **Right:** The use of the binary goal mask enables OVSegDT to search for the target object *stairs* more efficiently compared to DagRL, which takes extra steps in the environment and passes by the goal at the beginning of the episode.

(6) EALM, our proposed hybrid approach that jointly optimizes DAGger and PPO objectives. All experiments use the same transformer-based architecture with SigLIP features for RGB encoding and text goal representation, following the same architecture as in (Yokoyama et al. 2024b).

Table 2 confirms that naive PPO almost stalls. Success Rate (SR) alone, however, does not capture *how* the goal is reached. Table 2 complements the analysis with the average *collision count*. This metric, part of the PPO reward, is a convenient safety proxy: fewer collisions imply that the agent actually understands how to navigate rather than memorizing trajectories. DAGger attains near-perfect SR on the training scenes yet collides more than 50 times per unseen episode, illustrating severe over-fitting. The combined DAGger+PPO baseline shows lower collisions, but as the PPO

Method	Training steps	Val seen			Val unseen		
		SR↑	SPL↑	Collisions↓	SR↑	SPL↑	Collisions↓
PPO	200M	7.3	3.2	6.0	0.5	0.3	1.3
Dagger	200M	29.0	15.3	36.5	15.7	6.7	50.4
Dagger+PPO	200M	29.2	16.0	31.1	14.3	6.5	<u>35.8</u>
EarlySwitcher	200M	32.6	16.9	<u>25.6</u>	16.4	7.5	27.7
DagRL	300M	<u>41.3 ± 0.3</u>	<u>21.2 ± 0.3</u>	33.4	<u>18.3 ± 0.3</u>	<u>7.9 ± 0.1</u>	50.2
EALM (ours)	200M	42.5 ± 0.4	21.3 ± 0.2	25.3 ± 2.1	20.2 ± 0.5	8.8 ± 0.2	41.8 ± 2.5

Table 2: Quantitative comparison of switching strategies on HM3D-OVON benchmark. Besides classic navigation metrics (SR and SPL) we report mean collisions – a safety metric strongly correlated with generalisation. Lower values and bold highlights indicate better performance.

Method	EALM	Goal mask	\mathcal{L}_{seg}	SR↑	SPL↑	Collisions↓
DagRL	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	18.3±0.3	7.9 ± 0.1	50.2
OVSegDT	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	20.2±0.5	8.8 ± 0.2	41.8±2.5
OVSegDT	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<u>69.3±0.1</u>	40.8 ± 0.1	14.5 ± 0.4
OVSegDT	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	70.4 ± 0.2	<u>39.4±0.2</u>	<u>18.8±0.3</u>

Table 3: Comparison of methods using different observation types and training objectives on Val Unseen split of HM3D-OVON benchmark. Ground-truth segmentation masks are used as input.

Filtered mask finetune	Segmentation source	SR↑	SPL↑	Collisions↓
<input checked="" type="checkbox"/>	GT	70.4 ± 0.2	39.4 ± 0.2	18.8 ± 0.3
<input checked="" type="checkbox"/>	YOLOE	35.5 ± 0.2	16.4 ± 0.2	28.2 ± 0.03
<input checked="" type="checkbox"/>	YOLOE	35.1 ± 0.5	18.5 ± 0.3	28.8 ± 0.2
<input checked="" type="checkbox"/>	YOLOE+calib.	40.1 ± 1.1	20.9 ± 0.9	24.7 ± 3.1

Table 4: Analysis of different strategies of adaptation to predicted segmentation for OVSegDT method on Val Unseen split of HM3D-OVON benchmark.

term could not completely dominate the DAgger loss, the policy does not show improvement in the SR. EarlySwitcher demonstrates how delicate manual scheduling is. Switching the objectives linearly between 40 M and 60 M steps removes some collisions (down to ~ 26). Finding an “optimal” breakpoint would require costly hyper-parameter sweeps. EALM circumvents this dilemma by *continuously* adjusting the ratio between imitation and reinforcement in proportion to the policy entropy. When the agent is uncertain, behavior cloning dominates; as soon as it becomes confident, reinforcement learning takes over. This principled gating yields the best of both worlds: EALM cuts the sample complexity of training by 33% while achieving the lowest collision count on *val seen* and *val unseen* splits. In summary, collision count exposes over-fitting that success metrics alone may hide, and entropy-adaptive loss modulation proves crucial for producing policies that are *both* successful *and* safe. Additional training curves illustrating the performance over time for these strategies are provided in Appendix.

Training components. We analyze how each of our proposed components affects the performance of OVSegDT during training. In this experiment, we use ground-truth goal masks for the model variants that take a binary goal mask

as input. Table 3 shows that EALM improves overall navigation performance compared to the baseline training approach with a hard switch between DAgger and PPO objectives. Using the ground-truth goal mask significantly boosts navigation performance on unseen categories. The auxiliary segmentation loss increases the episode success rate but reduces the path efficiency (SPL) and increases the number of collisions. During reinforcement learning training, we do not optimize for path length. Also, we believe that in the open-vocabulary navigation task, the episode success rate is the most important metric; therefore, we retain this auxiliary training objective in subsequent experiments. Finding a more optimal integration of the segmentation and navigation functions for better optimization is left for future work.

Adaptation to predicted segmentation. We analyze how segmentation quality affects the navigation performance of our method and evaluate the contribution of our proposed techniques for improving navigation when using predicted segmentation. Table 4 shows that using the out-of-the-box YOLOE segmentation (Wang et al. 2025a) leads to a significant drop in navigation performance for unseen object categories. Fine-tuning on ground-truth masks filtered by size yields an improvement of +2.1% in SPL. Calibrating YOLOE confidence thresholds and vocabulary yields an additional gain of +4.6% in SR and +2.4% in SPL thanks to more accurate object category recognition. Overall, our method OVSegDT yields a gain of +21.8% in SR, +13% in SPL, and a 2× reduction in the average number of collisions compared to the baseline approach DagRL on the val unseen split of HM3D-OVON.

Conclusion

In this paper, we presented OVSegDT, a novel approach for open-vocabulary object navigation that integrates semantic segmentation and end-to-end learning. Our experiments demonstrate that our method significantly improves navigation performance, particularly on unseen object categories, by leveraging the synergy between segmentation and policy learning. The proposed unified training framework with a single-stage training pipeline switching between imitation and reinforcement training objectives further enhances learning efficiency. The limitations of the current approach include its dependence on the quality of the segmentation model and the lack of the ability to follow free-form text instructions. These directions are the subject of future work.

References

- Anderson, P.; Wu, Q.; Teney, D.; Bruce, J.; Johnson, M.; Sünderhauf, N.; Reid, I.; Gould, S.; and Van Den Hengel, A. 2018. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3674–3683.
- Cai, W.; Huang, S.; Cheng, G.; Long, Y.; Gao, P.; Sun, C.; and Dong, H. 2024. Bridging zero-shot object navigation and foundation models through pixel-guided navigation skill. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 5228–5234. IEEE.
- Chang, M.; Gervet, T.; Khanna, M.; Yenamandra, S.; Shah, D.; Min, S. Y.; Shah, K.; Paxton, C.; Gupta, S.; Batra, D.; et al. 2023. Goat: Go to any thing. *arXiv preprint arXiv:2311.06430*.
- Chen, J.; Lin, B.; Xu, R.; Chai, Z.; Liang, X.; and Wong, K.-Y. 2024. MapGPT: Map-Guided Prompting with Adaptive Path Planning for Vision-and-Language Navigation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 9796–9810.
- Cheng, A.-C.; Ji, Y.; Yang, Z.; Gongye, Z.; Zou, X.; Kautz, J.; Biyik, E.; Yin, H.; Liu, S.; and Wang, X. 2024. Navila: Legged robot vision-language-action model for navigation. *arXiv preprint arXiv:2412.04453*.
- Cheng, B.; Misra, I.; Schwing, A. G.; Kirillov, A.; and Girdhar, R. 2022. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1290–1299.
- Deitke, M.; VanderBilt, E.; Herrasti, A.; Weihs, L.; Ehsani, K.; Salvador, J.; Han, W.; Kolve, E.; Kembhavi, A.; and Mottaghi, R. 2022. ProcTHOR: Large-Scale Embodied AI Using Procedural Generation. *Advances in Neural Information Processing Systems*, 35: 5982–5994.
- Ehsani, K.; Gupta, T.; Hendrix, R.; Salvador, J.; Weihs, L.; Zeng, K.-H.; Singh, K. P.; Kim, Y.; Han, W.; Herrasti, A.; et al. 2023. Imitating shortest paths in simulation enables effective navigation and manipulation in the real world. *CoRR*.
- Hart, P. E.; Nilsson, N. J.; and Raphael, B. 1968. A formal basis for the heuristic determination of minimum cost paths. *IEEE transactions on Systems Science and Cybernetics*, 4(2): 100–107.
- He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.
- Kolve, E.; Mottaghi, R.; Han, W.; VanderBilt, E.; Weihs, L.; Herrasti, A.; Deitke, M.; Ehsani, K.; Gordon, D.; Zhu, Y.; et al. 2017. Ai2-thor: An interactive 3d environment for visual ai. *arXiv preprint arXiv:1712.05474*.
- Kulhánek, J.; Derner, E.; De Bruin, T.; and Babuška, R. 2019. Vision-based navigation using deep reinforcement learning. In *2019 european conference on mobile robots (ECMR)*, 1–8. IEEE.
- Lu, J.; Clark, C.; Lee, S.; Zhang, Z.; Khosla, S.; Marten, R.; Hoiem, D.; and Kembhavi, A. 2024. Unified-io 2: Scaling autoregressive multimodal models with vision language audio and action. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 26439–26455.
- Maksymets, O.; Cartillier, V.; Gokaslan, A.; Wijmans, E.; Galuba, W.; Lee, S.; and Batra, D. 2021. Thda: Treasure hunt data augmentation for semantic navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15374–15383.
- Radford, A.; Metz, L.; and Chintala, S. 2015. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*.
- Ramrakhya, R.; Batra, D.; Wijmans, E.; and Das, A. 2023. Pirlnav: Pretraining with imitation and rl finetuning for objectnav. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17896–17906.
- Santosh, D.; Achar, S.; and Jawahar, C. 2008. Autonomous image-based exploration for mobile robot navigation. In *2008 IEEE International Conference on Robotics and Automation*, 2717–2722. IEEE.
- Sethian, J. A. 1999. Fast marching methods. *SIAM review*, 41(2): 199–235.
- Staroverov, A.; Yudin, D. A.; Belkin, I.; Adeshkin, V.; Solomentsev, Y. K.; and Panov, A. I. 2020. Real-time object navigation with deep neural networks and hierarchical reinforcement learning. *IEEE Access*, 8: 195608–195621.
- Sun, X.; Liu, L.; Zhi, H.; Qiu, R.; and Liang, J. 2024. Prioritized semantic learning for zero-shot instance navigation. In *European Conference on Computer Vision*, 161–178. Springer.
- Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.
- Wang, A.; Liu, L.; Chen, H.; Lin, Z.; Han, J.; and Ding, G. 2025a. Yoloe: Real-time seeing anything. *arXiv preprint arXiv:2503.07465*.
- Wang, S.; Wang, Y.; Li, W.; Cai, X.; Wang, Y.; Chen, M.; Wang, K.; Su, Z.; Li, D.; and Fan, Z. 2025b. Aux-Think: Exploring Reasoning Strategies for Data-Efficient Vision-Language Navigation. *arXiv preprint arXiv:2505.11886*.
- Wijmans, E.; Essa, I.; and Batra, D. 2022. Ver: Scaling on-policy rl leads to the emergence of navigation in embodied rearrangement. *Advances in Neural Information Processing Systems*, 35: 7727–7740.
- Wijmans, E.; Kadian, A.; Morcos, A.; Lee, S.; Essa, I.; Parikh, D.; Savva, M.; and Batra, D. 2020. DD-PPO: Learning Near-Perfect PointGoal Navigators from 2.5 Billion Frames. In *International Conference on Learning Representations*.
- Wu, H.; Jing, Y.; Cheang, C.; Chen, G.; Xu, J.; Li, X.; Liu, M.; Li, H.; and Kong, T. 2024. Unleashing Large-Scale Video Generative Pre-training for Visual Robot Manipulation. In *ICLR*.

Yadav, K.; Ramrakhyta, R.; Ramakrishnan, S. K.; Gervet, T.;

Turner, J.; Gokaslan, A.; Maestre, N.; Chang, A. X.; Batra, D.; Savva, M.; et al. 2023. Habitat-matterport 3d semantics dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4927–4936.

Yin, H.; Xu, X.; Wu, Z.; Zhou, J.; and Lu, J. 2024. Sg-nav: Online 3d scene graph prompting for llm-based zero-shot object navigation. *Advances in neural information processing systems*, 37: 5285–5307.

Yokoyama, N.; Ha, S.; Batra, D.; Wang, J.; and Bucher, B. 2024a. Vlfm: Vision-language frontier maps for zero-shot semantic navigation. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 42–48. IEEE.

Yokoyama, N.; Ramrakhyta, R.; Das, A.; Batra, D.; and Ha, S. 2024b. Hm3d-ovon: A dataset and benchmark for open-vocabulary object goal navigation. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 5543–5550. IEEE.

Zeng, K.-H.; Zhang, Z.; Ehsani, K.; Hendrix, R.; Salvador, J.; Herrasti, A.; Girshick, R.; Kembhavi, A.; and Weihs, L. 2025. PoliFormer: Scaling On-Policy RL with Transformers Results in Masterful Navigators. In *Conference on Robot Learning*, 408–432. PMLR.

Zhai, X.; Mustafa, B.; Kolesnikov, A.; and Beyer, L. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, 11975–11986.

Zhang, J.; Wang, K.; Wang, S.; Li, M.; Liu, H.; Wei, S.; Wang, Z.; Zhang, Z.; and Wang, H. 2024. Uni-navid: A video-based vision-language-action model for unifying embodied navigation tasks. *arXiv preprint arXiv:2412.06224*.

Zhen, H.; Qiu, X.; Chen, P.; Yang, J.; Yan, X.; Du, Y.; Hong, Y.; and Gan, C. 2024. 3D-VLA: a 3D vision-language-action generative world model. In *Proceedings of the 41st International Conference on Machine Learning*, 61229–61245.

Zhou, G.; Hong, Y.; and Wu, Q. 2024. Navgpt: Explicit reasoning in vision-and-language navigation with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 7641–7649.

Zhou, K.; Zheng, K.; Pryor, C.; Shen, Y.; Jin, H.; Getoor, L.; and Wang, X. E. 2023. Esc: Exploration with soft common-sense constraints for zero-shot object navigation. In *International Conference on Machine Learning*, 42829–42842. PMLR.

Zhu, Z.; Wang, X.; Li, Y.; Zhang, Z.; Ma, X.; Chen, Y.; Jia, B.; Liang, W.; Yu, Q.; Deng, Z.; et al. 2025. Move to Understand a 3D Scene: Bridging Visual Grounding and Exploration for Efficient and Versatile Embodied Navigation. *arXiv preprint arXiv:2507.04047*.

Ziliotto, F.; Campari, T.; Serafini, L.; and Ballan, L. 2025. TANGO: training-free embodied AI agents for open-world tasks. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 24603–24613.

Reproducibility Checklist

Instructions for Authors:

This document outlines key aspects for assessing reproducibility. Please provide your input by editing this .tex file directly.

For each question (that applies), replace the “Type your response here” text with your answer.

Example: If a question appears as

```
\question{Proofs of all novel claims  
are included} { (yes/partial/no) }  
Type your response here
```

you would change it to:

```
\question{Proofs of all novel claims  
are included} { (yes/partial/no) }  
yes
```

Please make sure to:

- Replace ONLY the “Type your response here” text and nothing else.
- Use one of the options listed for that question (e.g., yes, no, partial, or NA).
- **Not** modify any other part of the \question command or any other lines in this document.

You can \input this .tex file right before \end{document} of your main file or compile it as a stand-alone document. Check the instructions on your conference’s website to see if you will be asked to provide this checklist with your paper or separately.

1. General Paper Structure

- 1.1. Includes a conceptual outline and/or pseudocode description of AI methods introduced (yes/partial/no/NA) **yes**
- 1.2. Clearly delineates statements that are opinions, hypothesis, and speculation from objective facts and results (yes/no) **yes**
- 1.3. Provides well-marked pedagogical references for less-familiar readers to gain background necessary to replicate the paper (yes/no) **yes**

2. Theoretical Contributions

- 2.1. Does this paper make theoretical contributions? (yes/no) **no**
If yes, please address the following points:
 - 2.2. All assumptions and restrictions are stated clearly and formally (yes/partial/no) **Type your response here**

- 2.3. All novel claims are stated formally (e.g., in theorem statements) (yes/partial/no) [Type your response here](#)
- 2.4. Proofs of all novel claims are included (yes/partial/no) [Type your response here](#)
- 2.5. Proof sketches or intuitions are given for complex and/or novel results (yes/partial/no) [Type your response here](#)
- 2.6. Appropriate citations to theoretical tools used are given (yes/partial/no) [Type your response here](#)
- 2.7. All theoretical claims are demonstrated empirically to hold (yes/partial/no/NA) [Type your response here](#)
- 2.8. All experimental code used to eliminate or disprove claims is included (yes/no/NA) [Type your response here](#)

3. Dataset Usage

- 3.1. Does this paper rely on one or more datasets? (yes/no) **yes**

If yes, please address the following points:

- 3.2. A motivation is given for why the experiments are conducted on the selected datasets (yes/partial/no/NA) **yes**
- 3.3. All novel datasets introduced in this paper are included in a data appendix (yes/partial/no/NA) **NA**
- 3.4. All novel datasets introduced in this paper will be made publicly available upon publication of the paper with a license that allows free usage for research purposes (yes/partial/no/NA) **NA**
- 3.5. All datasets drawn from the existing literature (potentially including authors' own previously published work) are accompanied by appropriate citations (yes/no/NA) **yes**
- 3.6. All datasets drawn from the existing literature (potentially including authors' own previously published work) are publicly available (yes/partial/no/NA) **yes**
- 3.7. All datasets that are not publicly available are described in detail, with explanation why publicly available alternatives are not scientifically satisfying (yes/partial/no/NA) **NA**

4. Computational Experiments

- 4.1. Does this paper include computational experiments? (yes/no) **yes**

If yes, please address the following points:

- 4.2. This paper states the number and range of values tried per (hyper-) parameter during development of

the paper, along with the criterion used for selecting the final parameter setting (yes/partial/no/NA) **yes**

- 4.3. Any code required for pre-processing data is included in the appendix (yes/partial/no) **yes**
- 4.4. All source code required for conducting and analyzing the experiments is included in a code appendix (yes/partial/no) **partial**
- 4.5. All source code required for conducting and analyzing the experiments will be made publicly available upon publication of the paper with a license that allows free usage for research purposes (yes/partial/no) **yes**
- 4.6. All source code implementing new methods have comments detailing the implementation, with references to the paper where each step comes from (yes/partial/no) **yes**
- 4.7. If an algorithm depends on randomness, then the method used for setting seeds is described in a way sufficient to allow replication of results (yes/partial/no/NA) **yes**
- 4.8. This paper specifies the computing infrastructure used for running experiments (hardware and software), including GPU/CPU models; amount of memory; operating system; names and versions of relevant software libraries and frameworks (yes/partial/no) **yes**
- 4.9. This paper formally describes evaluation metrics used and explains the motivation for choosing these metrics (yes/partial/no) **yes**
- 4.10. This paper states the number of algorithm runs used to compute each reported result (yes/no) **yes**
- 4.11. Analysis of experiments goes beyond single-dimensional summaries of performance (e.g., average; median) to include measures of variation, confidence, or other distributional information (yes/no) **yes**
- 4.12. The significance of any improvement or decrease in performance is judged using appropriate statistical tests (e.g., Wilcoxon signed-rank) (yes/partial/no) **partial**
- 4.13. This paper lists all final (hyper-)parameters used for each model/algorithm in the paper's experiments (yes/partial/no/NA) **yes**

Appendix

Reproducibility. All experiments were conducted on a machine equipped with two NVIDIA A100 GPUs and an Intel Xeon Gold 6338 CPU, running Ubuntu 22.04 with 96 GB of RAM. We also provide code for reproducing the results in the Code Appendix.

1 Hyperparameters

Table 5 lists the key hyperparameters used for the policy model.

Table 5: Policy Model Hyperparameters

Component	Parameter	Value
Visual encoder	Backbone Network	SigLIP
	Fusion Type	Concatenation
	Use Visual Query	True
	Use Residual Connections	True
Transformer	Base Model	LLaMA
	Number of Layers	4
	Number of Attention Heads	8
	Hidden Dimension Size	512
	MLP Hidden Dimension Size	1024
	Max Context Length	100
	Shuffle Position IDs	True
Training parameters	Learning Rate	2.5×10^{-4}
	Parallel environments	40

Table 6 lists the key hyperparameters used for the Proximal Policy Optimization (PPO) algorithm.

Table 7 details the configuration parameters for the agent and its observations within the simulation environment.

2 Analysis of Training Strategies

As shown in Figure 4, the naïve PPO agent makes little progress. However, Success Rate (SR) alone doesn’t reveal

how the agent reaches its goal. To address this, Figure 5 includes the average number of *collisions*, a useful proxy for safety since it’s directly tied to the PPO reward—fewer collisions suggest the agent is genuinely learning to navigate rather than memorizing paths. While DAgger achieves near-perfect SR on training scenes, it averages over 50 collisions per episode, indicating strong overfitting. The DAgger+PPO combined loss reduces collisions initially, but because the PPO signal doesn’t fully outweigh the DAgger loss, the agent regresses to unsafe behavior, leading to a drop in validation performance.

Table 6: Proximal Policy Optimization (PPO) Hyperparameters

Component	Parameter	Value
PPO	Clip Parameter (ϵ)	0.2
	PPO Epochs	1
	Mini-batches	2
	Value Loss Coefficient	0.5
	Entropy Coefficient	0.01
	Learning Rate (LR)	2.5×10^{-4}
	Adam Epsilon (ϵ)	1×10^{-5}
	Max Gradient Norm	0.2
	Steps per Update	100
	Use GAE	True
	Discount Factor (γ)	0.99
	GAE Lambda (λ / τ)	0.95
	Linear Clip Decay	False
	Linear LR Decay	True
	Reward Window Size	50
	Normalized Advantage	False

Table 7: Agent and Observation Configuration Parameters

Component	Parameter	Value
NavMesh	Agent Max Climb	0.1
	Cell Height	0.05
Environment	Turn Angle (degrees)	30
	Height	1.41
Agent	Radius	0.17
	RGB Sensor Width (pixels)	360
	RGB Sensor Height (pixels)	640
	Horizontal FOV (degrees)	42
	RGB Sensor Position (x, y, z)	[0, 1.31, 0]

The EarlySwitcher highlights the sensitivity of manual objective scheduling. Gradually shifting from imitation to reinforcement between 40M and 60M steps reduces collisions (to about ~ 26), but it takes over 150M steps for the policy to regain the SR lost during the transition. Identifying the optimal switching point would require extensive hyperparameter tuning. EALM avoids this issue by *continuously* adjusting the balance between imitation learning and reinforcement learning based on policy entropy. When the agent is uncertain, imitation dominates; as confidence grows, reinforcement naturally takes over. This adaptive approach combines the strengths of both methods—EALM converges

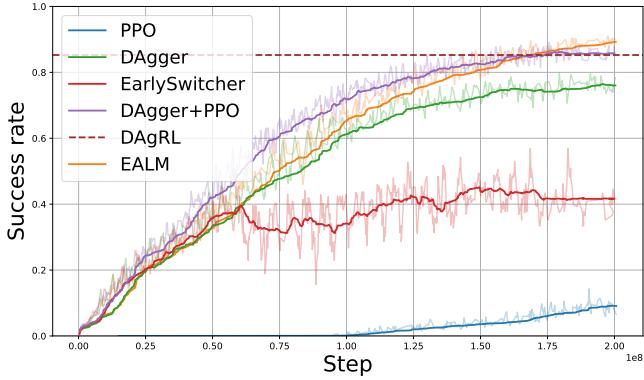


Figure 4: Training curves of **Success Rate** (higher is better) for the considered switching strategies. Our entropy-adaptive EALM reaches top performance with the fewest samples.

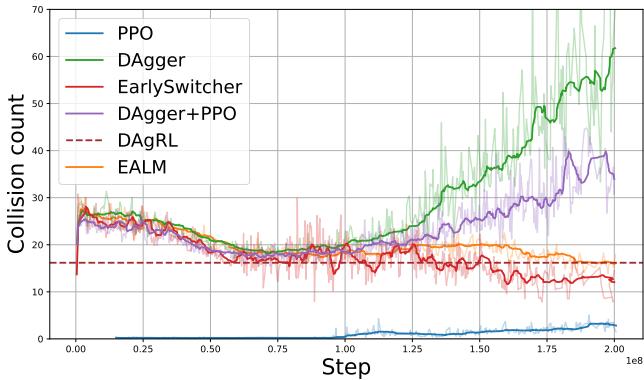


Figure 5: Training curves of **Collision Count** (lower is better). Collisions serve as a safety proxy; EALM steadily reduces collisions whereas other methods eventually over-fit and regress.

twice as fast as the best baseline while achieving the fewest collisions.

3 Analysis of the confidence threshold of the segmentation model on SR OVSegDT

In our experiments with predicted segmentation, we use a pretrained YOLOE model based on YOLOv8-L. We analyze how the confidence threshold of the YOLOE model for each category affects the percentage of episodes with successful navigation to objects of that category. Figures 6, 7, and 10 illustrate the relationship between success rate and different confidence thresholds for the top 10 categories by number of episodes from each split of HM3D-OVON.

The analysis shows that some categories are challenging for the pretrained YOLOE model and require a low confidence threshold for successful recognition (e.g., *plant* from *val unseen* or *armchair* from *val seen synonyms*). On the other hand, some categories require a high confidence threshold to be reliably distinguished from others (e.g., *kitchen counter* from *val seen synonyms* or *mirror* from *val*

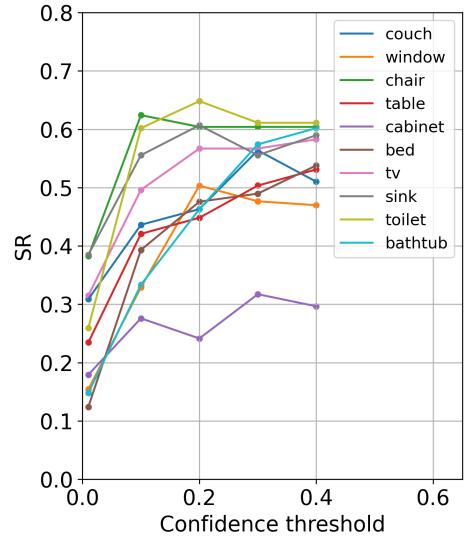


Figure 6: Dependence of the success rate (SR) on the confidence threshold of the YOLOE model for different categories from the *val seen* split of HM3D-OVON. The plots are shown for the top 10 categories by number of episodes.

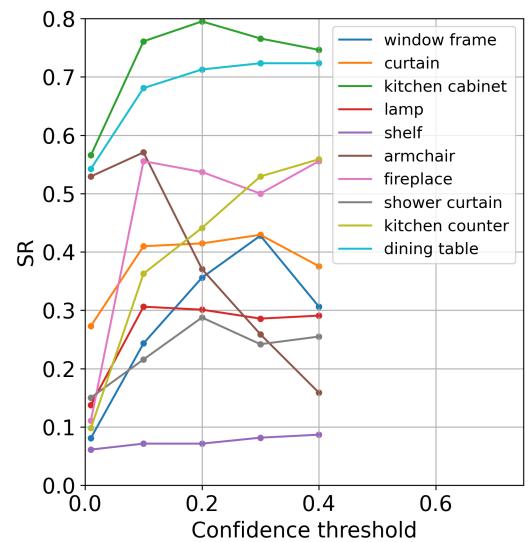


Figure 7: Dependence of the success rate (SR) on the confidence threshold of the YOLOE model for different categories from the *val seen synonyms* split of HM3D-OVON. The plots are shown for the top 10 categories by number of episodes.

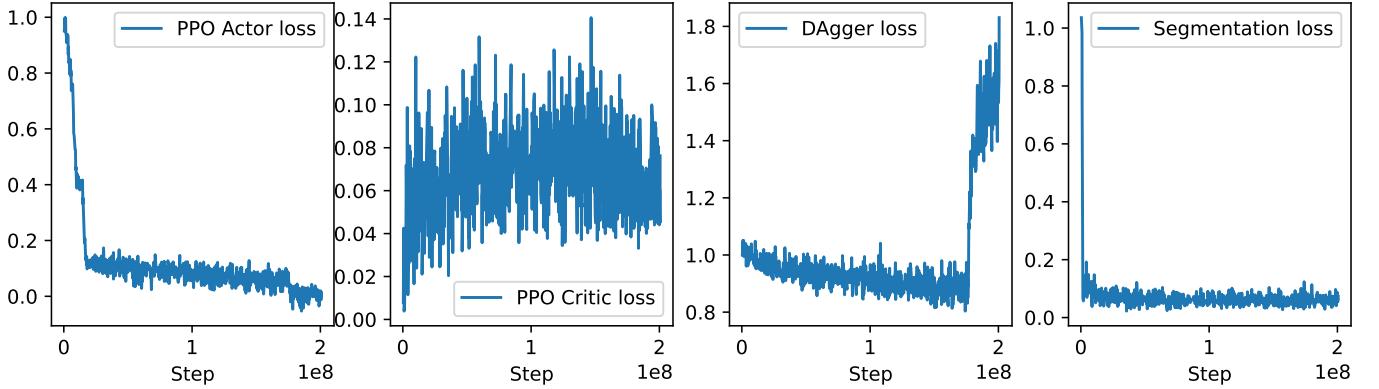


Figure 8: Training losses for OVSegDT over 200 M environment steps. **Left to right:** (i) PPO actor loss; (ii) PPO critic loss; (iii) DAgger (behaviour-cloning) loss; (iv) Segmentation loss.

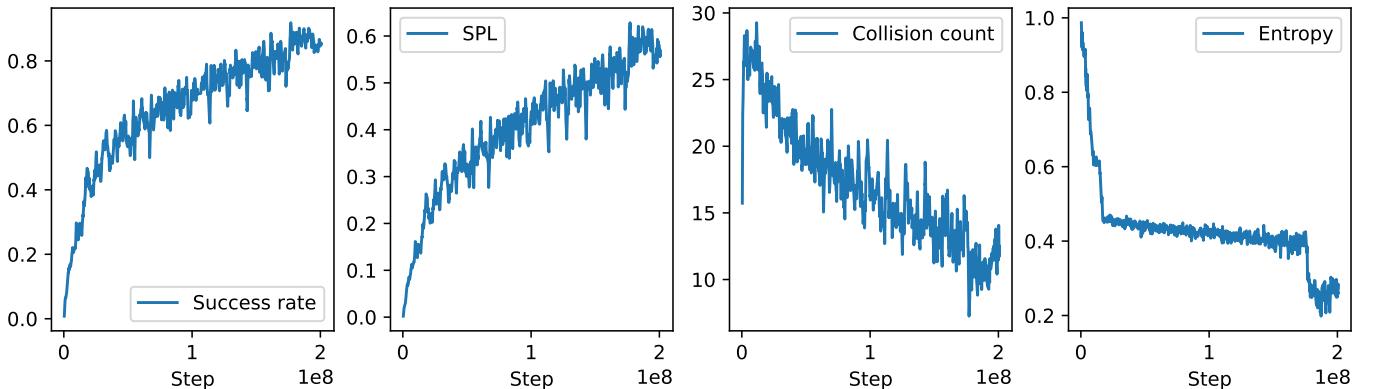


Figure 9: Metrics on the HM3D-OVON *train* split during training. **Left to right:** (i) Success Rate (SR); (ii) Success weighted by Path Length (SPL); (iii) Collision count; (iv) Policy entropy.

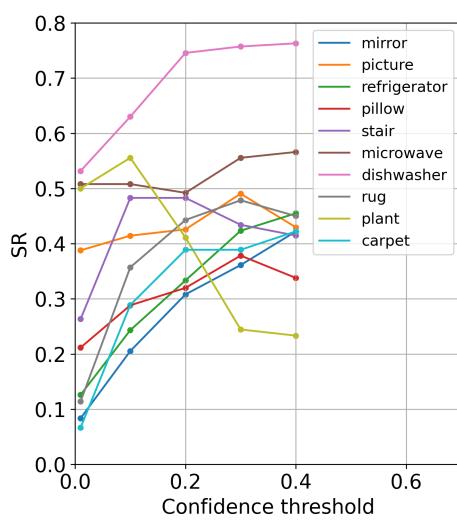


Figure 10: Dependence of the success rate (SR) on the confidence threshold of the YOLOE model for different categories from the *val unseen* split of HM3D-OVON. The plots are shown for the top 10 categories by number of episodes.

unseen). For most categories, the optimal confidence threshold falls within the range of 0.1 to 0.3.

4 Analysis of Training Loss Components

Figure 8 visualizes how each loss term evolves over 200 M environment steps. The *PPO actor* loss drops within the first 10 M steps and then asymptotically approaches zero, reflecting rapid policy improvement. The *critic* loss spikes early, stabilizing once value estimates become consistent. During the imitation-heavy phase the *DAgger* loss is low; it rises after ~ 160 M steps when EALM hands full control to PPO, confirming that the optimization objective has indeed shifted. Finally, the *segmentation* loss falls quickly and remains an order of magnitude lower than the policy losses, indicating that the auxiliary head is able to reconstruct accurate goal masks throughout training.

Performance trends on the HM3D-OVON *train* split are shown in Figure 9. Both **Success Rate** and **SPL** increase monotonically, while the **collision count** falls to fewer than ten per episode - evidence that the agent learns safer, shorter paths as training progresses. The **entropy** curve drops sharply during the first imitation-dominated stage, slowly decreases while EALM mixes objectives, and exhibits a second rough decline when the algorithm switches

H ₀ Hypothesis	Metric	Split	p-value	Conclusion
No significant evidence that DAgger+PPO outperforms DAgger.	collisions	val seen	0.004	Significant difference
No significant evidence that DAgger+PPO outperforms DAgger.	collisions	val unseen	0.03	Significant difference
No significant evidence that EarlySwitcher outperforms DAgger+PPO.	collisions	val seen	0.008	Significant difference
No significant evidence that EarlySwitcher outperforms DAgger+PPO.	collisions	val unseen	0.01	Significant difference
No significant evidence that EALM outperforms DagRL.	SR	val seen	0.001	Significant difference
No significant evidence that EarlySwitcher outperforms DAgger+PPO.	SR	val unseen	0.03	Significant difference
No significant evidence that EarlySwitcher outperforms DAgger+PPO.	SPL	val unseen	0.002	Significant difference

Table 8: Statistical comparison of different switching strategies on HM3D-OVON benchmark.

H ₀ Hypothesis	Metric	Split	p-value	Conclusion
No significant evidence that OVSegDT+Goal Mask outperforms OVSegDT.	SR	val unseen	5e-5	Significant difference
No significant evidence that OVSegDT+Goal Mask outperforms OVSegDT.	SPL	val unseen	2e-6	Significant difference
No significant evidence that OVSegDT+Goal Mask+ \mathcal{L}_{seg} outperforms OVSegDT+Goal Mask.	SR	val unseen	0.005	Significant difference

Table 9: Statistical comparison of different observation types and training objectives on Val Unseen split of HM3D-OVON benchmark.

H ₀ Hypothesis	Metric	Split	p-value	Conclusion
No significant evidence that OVSegDT+Filtered mask finetune outperforms OVSegDT.	SPL	val unseen	0.008	Significant difference
No significant evidence that OVSegDT+Filtered mask finetune+YOLOE calibrated outperforms OVSegDT+Filtered mask finetune.	SR	val unseen	0.002	Significant difference
No significant evidence that OVSegDT+Filtered mask finetune+YOLOE calibrated outperforms OVSegDT+Filtered mask finetune.	SPL	val unseen	0.009	Significant difference

Table 10: Statistical comparison of different strategies of adaptation to predicted segmentation for OVSegDT method on Val Unseen split of HM3D-OVON benchmark.

to a PPO-only loss near the end of training. Importantly, this late entropy reduction *does not* introduces instability: all task metrics continue to improve smoothly, demonstrating that EALM’s automatic transition preserves training stability and leads to a confident yet robust final policy.

5 Statistical analysis

We analyze the statistical significance of the impact of different components of our method. Each algorithm is run three times, and we perform the one-sided t-test for unpaired samples to determine the level of statistical significance. Tables 8, 9, and 10 present the comparison results and corresponding conclusions. Thus, the conclusions presented in the main text are supported by the statistical significance of the results.

6 Analysis of observation components impact on navigation performance

The OVSegDT method uses two types of goal prompts at inference time: a textual instruction and a visual cue in the form of a binary mask of the target object. We analyze the impact of each of these components during inference in Table 11. The observation setting Goal Mask+Text Goal corresponds to the original version of OVSegDT. When using only the Goal Mask at inference time, we replace the part of the observation corresponding to the text instruction with a dummy input consisting of the embedding of the “table” category from the val seen split. In this case, we observe a slight drop in navigation performance. In the case of navigation using only the text goal, we always feed a zero

mask to the model during inference. This leads to a complete breakdown of OVSegDT’s navigation, highlighting the importance of high-quality masks for successful navigation. Thus, although binary goal masks are a key component of the observation, the textual instruction also contributes to improving navigation quality in OVSegDT.

Method	Observation	SR	SPL	Collisions
OVSegDT Goal Mask+Text Goal		70.4 \pm 0.2	39.4 \pm 0.2	18.8 \pm 0.3
OVSegDT Goal Mask		65.3 \pm 0.3	34.2 \pm 0.2	20.0 \pm 0.2
OVSegDT Text Goal		0.1 \pm 0.08	0.0 \pm 0.0	40.8 \pm 0.5

Table 11: Analysis of the impact of the goal mask and textual instruction on navigation performance on Val Unseen of HM3D-OVON benchmark.